“Thinking Statistically”: probability& statistics taught with heavy emphasis on using simulations and re-sampling methods to both DO statistics (analyze data) and, more importantly, understand concepts

Three motivations:

1. Most existing statistics courses are simultaneously too superficial (e.g. flow-chart approach to test selection) and too detailed (e.g. calculations of covariance), and fail to develop intuitions that are central to statistical thinking (e.g. What is a p-value? a Bayes Factor?).
2. There is an abundance of good web-based material that can be taken advantage of if the students understand basic concepts and have basic programming skills.
3. We are not teaching our students the statistical approaches they will actually need in the world of big data. c.f. Efron, “Large-scale inference.”

Quarter course: 18 contact hours

Question: How do we factor in time watching video-taped lectures?

Put another way, what should be the ratio of class time to preparation time?

According to the University of Iowa, it should be about 1:2. Can we also use this formula for a flipped classroom? Which part should be considered “prep” and which part “in class”? [Note: this ratio confirmed by multiple curriculum fellows (Johanna et al.). I still worry about 4 hours of video / week.]

Approach: Prior to class sessions, students will view 4 hours of lectures on-line from Brian Healy’s course. In class, we will focus on demos to build intuition and practice different approaches to analyzing real data sets. All in-class programming will be done in MATLAB.

**Syllabus**

Week #1: Thinking about data: tidy data/clean data; data visualization; 40k-ft view

Video: 1: Introduction to Biostatistics; 2: Basics of Probability

In class: distribution plots vs. box plots; PIN data visualization; dual-code for image analysis

In class: Bayes vs. Frequentist; Bayes dice demo; central limit theorem demo; James-Stein demo

Reading: Leek, “How to share data with a statistician”; Wickham, “Tidy Data”; Allen et al., Neuron 2012

Intro (15):

* Why this course?
* Importance of becoming a life-long statistician: e.g. subscribe to datacolada.org
* Course mechanics and expectations

Data (45)

* How to share data with a statistician. tidy data / clean data
* View your data in the rawest form that is visually comprehensible: PIN data
* Think hard about effective ways to visualize your data: PIN data cont’d
* In publications: Hide less; show more:
  + distribution plots (a.k.a. ‘violin’ plots) vs. whisker plots vs. bar plots
  + matrix of covariate plots
  + dual-code colorbars for images

Probability (60)

* using simulations to solve (difficult!) probability problems
  + class picks number ‘randomly’ [1,4]
  + extension to birthday problem
* using simulations to gain intuition
  + CLT demo
* Frequentist vs. Bayes from 40k feet
  + Bayes theorem: 4 views
    - rule for inverting conditional probabilities
    - method for combining new data with existing knowledge
      * twin problem: U/S shows 2 boys; 1/3 twins identical; P(twins/US)?
    - method for updating beliefs as evidence accumulates over time
      * Bayes dice demo (Bayes factors, etc.)
    - a religion (compare and contrast w/ frequentist)
  + Converting p-values to Bayes factors

Week #2: Thinking about data: “torturing data ethically” (XLM): Power, P-hacking and Reproducibility

Video: 3. Hypothesis testing/t-test; 4. One-way ANOVA

In class: multiple comparisons; “researcher degrees of freedom”; simulation of stopping rule for data collection; distribution of p-values under H0; P-curve; power calculations via simulation

Reading: Simmons et al., False Positive Statistics 2011; Simonsohn et al. p-Curve and effect size, 2014

Other resources:

<http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/>

Button et al. “Power failure”, Nat. Rev. Neurosci. 2013

Paul Meehl lectures: <http://meehl.umn.edu/recordings/philosophical-psychology-1989>

Gelman & Loken (2014) “Forking paths”

Week #4: An Introduction to the bootstrap

Video: 5. Nonparametric approaches; 6. Analysis of proportions

In class: bootstrapping confidence intervals; hypothesis testing; power calculations via simulation

Reading: Efron & Tibshirani

Week #5: Linear Regression I

Video: 7. Linear regression and correlation; 12. Regression diagnostics

In class: over-fitting; cross-validation; regularization

Reading: “Applied Regression Analysis and Other Multivariable Methods” by Kleinbaum, Kupper, Nizam & Muller

Week #6: Linear Regression II

Video: 11. Multiple Linear Regression I; 16/17. Logistic Regression

In class: Using the GLM to do logistic regression on microstim experiment data

Reading: “Regression Modeling Strategies” by Harrell

Week #7: Thinking about data: dimensionality reduction

Video: https://www.youtube.com/watch?v=\_UVHneBUBW0

In class: spike sorting; Novembre et al. 2008?; Stephen Holtz & Alex Batchelor exercise: PCA neurons

Reading: Shlens tutorial on PCA

Week #8: Introduction to Machine Learning

Video: Jan???

In class: gradient descent using an objective function; regression as a classifier; SVM

General Resources:

Datacolada: <http://datacolada.org/>

[MIT 18.05](https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/index.htm)

Nature’s statistics for biologists: <http://www.nature.com/collections/qghhqm/content/practical-guides>